Available online at www.sciencedirect.com
SCIENCE DIRECT
Gene 346 (2005) 215–219
GENE
SECTION
EVOLUTIONARY GENOMICS
www.elsevier.com/locate/gene

Eighty percent of proteins are different between humans and chimpanzees

Galina Glazko, Vamsi Veeramachaneni, Masatoshi Nei, Wojciech Makałowski*

Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, PA 16802, USA
Department of Biology, Pennsylvania State University, University Park, PA 16802, USA

Received 12 August 2004; received in revised form 1 October 2004; accepted 5 November 2004
Available online 29 January 2005
Received by T. Gojobori

Abstract

The chimpanzee is our closest living relative. The morphological differences between the two species are so large that there is no problem in distinguishing between them. However, the nucleotide difference between the two species is surprisingly small. The early genome comparison by DNA hybridization techniques suggested a nucleotide difference of 1–2%. Recently, direct nucleotide sequencing confirmed this estimate. These findings generated the common belief that the human is extremely close to the chimpanzee at the genetic level. However, if one looks at proteins, which are mainly responsible for phenotypic differences, the picture is quite different, and about 80% of proteins are different between the two species. Still, the number of proteins responsible for the phenotypic differences may be smaller since not all genes are directly responsible for phenotypic characters.

Keywords: Human; Chimpanzee; Genetic distance; Protein identity; Nucleotide identity

1. Introduction

In terms of nucleotide differences, the human is closer to the chimpanzee than to any other hominoid species. The early genome comparison by DNA hybridization suggested a nucleotide difference of 1–2% (Kohne, 1970; Sibley and Ahlquist, 1984). Recently, direct nucleotide sequencing confirmed this estimate (Goodman, 1995; Chen and Li, 2001; Ebersberger et al., 2002; Watanabe et al., 2004).

However, a large portion (about 98%) of the human genome is known to be non-protein-coding DNA, and the estimate of 1–2% nucleotide difference is largely based on the comparison of non-protein-coding DNA, which has little effect on phenotypic characters. Therefore, for the general public who are interested in phenotypic differences, this is clearly misleading. A better way of measuring the genetic difference is to consider functional genes or proteins as the units of comparison, because these are the genetic units that control phenotypic characters. To do this, we compiled 127 human and chimp orthologous proteins (44,000 amino acid residues) from GenBank. Only 25 (20%) of these proteins showed the identical amino acid sequence between humans and chimpanzees. In other words, the proportion of different proteins was 80%, in contrast to the 1–2% difference at the nucleotide level. How these differences are related to the morphological differences is unclear at present, but it is quite possible that a large proportion of phenotypic differences are caused by a relatively small number of

Abbreviations: PAM, percent of accepted mutations; MHC, major histocompability complex; d_s, synonymous substitution distance; d_n, nonsynonymous substitution distance; PBL, Pamilo–Bianki–Li method; mNG, modified Nei-Gojobori method; NCBI, National Center for Biotechnology Information; GO, Gene Ontology.

* Corresponding author. Department of Biology, Pennsylvania State University, 514 Muelle Lab, University Park, PA 16802, USA.
E-mail address: wojtek@psu.edu (W. Makałowski).

0378-1119/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2004.11.003

regulatory mutations (King and Wilson, 1975) or major effect genes (Nei, 1987).

2. Materials and methods

2.1. Apes protein sequences

All human protein sequences known as of March 2003 were compared with all ape proteins available at that time. The human and ape protein data set was prepared as follows. First, all the ape proteins were downloaded from GenBank, and then all the identical proteins and those that are substrings of other proteins were merged. Finally, since we were interested only in full-length proteins, we checked all the sequences for the following criteria: length of at least 10 amino acids with the initiation codon of methionine. It resulted in the following data set—human: 71,334; chimpanzee: 384; gorilla: 157; orangutan: 152; and gibbon: 82 proteins.

2.2. Assignment of orthology relationships

To find ape orthologs of human proteins, the BLAST search of each human protein sequences against all ape sequences was performed with the following parameters: E-value cut-off (*e*), 0.001; matrix (*M*), PAM30; gap extension penalty (*E*), 1; gap opening penalty (*G*), 9; number of database sequences to show one-line description for (*v*), 10; number of database sequences to show alignment for (*b*), 10; and low complexity filtering for lookup table only (*F*), "m L". The resulting alignments were manually analyzed. Multigene families such as major histocompability complex (MHC), immunoglobulin, olfactory receptor, and KIR receptor gene families were excluded from the analysis because of difficulties in detecting orthologous relationships. Mitochondrial proteins were also excluded from the analysis. The final ortholog data set consisted of 137 human, 127 chimpanzee, 60 gorilla, 56 orangutan, and 31 gibbon sequences. Here, some human genes were orthologous only to some ape genes.

2.3. Statistical analysis

Orthologous sequence pairs were transformed into 137 multiple species orthology groups using the single linkage approach, and multiple sequence alignments were obtained by using ClustalW with default parameters. Using these alignments, we computed the number of amino acid and nucleotide differences per site. We also computed the synonymous (d_S) and nonsynonymous (d_N) nucleotide substitutions per site using the modified Nei–Gojobori (mNG) and the Pamilo–Bianki–Li (PBL) methods (see Nei and Kumar, 2000). The statistical test (*Z* test) of the difference $d_N - d_S$ was conducted by computing the standard error of $d_N - d_S$ analytically or with the boostrap test. All these computations were done by using the computer program MEGA2 (Kumar et al., 2001). We did not use the Goldman–Yang method, because the statistical test of $d_N - d_S$ was not available in the program PAML (Yang, 2003).

2.4. Database of orthologous genes

All the results are stored in a *mysql* database. With a simple web interface, readers can have access to the following information: lists of identical protein sequences between hominoid species pairs, identical coding DNA sequences between them, and the proteins that do not appear in ortholog sets. We also provide a link to a table with orthologous gene groups from which multiple amino acid sequence alignment is available, along with links to original sequence information via Entrez system at the National Center for Biotechnology Information (NCBI). Finally, there is a simple search tool that enables a keyword search based on information included in definition line of the original records. The database is accessible at http://warta.bio.psu.edu/ApesDB/.

3. Results and discussion

3.1. Statistical properties of the data

Using the orthologs specified in Section 2.2, we selected 411 hominoid orthologous genes clustered in 137 orthology groups. The number of orthologous genes for each species pair is summarized in Table 1. Five-species orthology groups were established only for 18 proteins. The data set used here appears to be a random set of proteins, because the size distribution of the proteins was similar to that of the entire set of human proteins (see Fig. 1). The list of all proteins used in this study is provided on the project web site http://warta.bio.psu.edu/ApesDB/. Only 25 out of the 127 chimpanzee proteins (20%) were identical to their human orthologs (see Table 2 for the list of these proteins). This is approximately in accordance with a random distribution of the 0.6% nonsynonymous substitutions across proteins of average length of 330 amino acids (V.V. and W.M., unpublished simulation results). Interestingly, there are several genes that showed the identical nucleotide

Table 1
Number of orthologous gene pairs used in this study

	Human	Chimpanzee	Gorilla	Orangutan
Chimpanzee	127			
Gorilla	60	56		
Orangutan	56	52	44	
Gibbon	31	26	22	24